Research Article | Vol. 12, No. 12 / 1 Dec 2021 / Biomedical Optics Express 7568
Biomedical Optics EXPRESS

Blood identification at the single-cell level based on a combination of laser tweezers Raman spectroscopy and machine learning

ZIQI WANG,[1] YIMING LIU,[1] WEILAI LU,[2] YU VINCENT FU,[2,3] AND ZHEHAI ZHOU[1,4]

[1]Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instruments, Beijing Information Science and Technology University, Beijing, China
[2]State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China
[3]fuyu@im.ac.cn
[4]zhouzhehai@bistu.edu.cn

Abstract: Laser tweezers Raman spectroscopy (LTRS) combines optical tweezers technology and Raman spectroscopy to obtain biomolecular compositional information from a single cell without invasion or destruction, so it can be used to "fingerprint" substances to characterize numerous types of biological cell samples. In the current study, LTRS was combined with two machine learning algorithms, principal component analysis (PCA)-linear discriminant analysis (LDA) and random forest, to achieve high-precision multi-species blood classification at the single-cell level. The accuracies of the two classification models were 96.60% and 96.84%, respectively. Meanwhile, compared with PCA-LDA and other classification algorithms, the random forest algorithm is proved to have significant advantages, which can directly explain the importance of spectral features at the molecular level.

1. Introduction

Blood identification from different species is of great significance in many areas. For example, the quarantine of imported and exported animal products under the current epidemic environment, forensic analysis in criminal cases, and the protection of endangered wildlife are all heavily dependent on blood-based species identification. Many analytical methods have been used to identify blood from different species. Traditional methods include immunochromatographic assays, redox reactions, and enzyme immunoassays, among others [1–3]. With progress in modern instrumentation and analytical methods, emerging technologies such as high-performance liquid chromatography [4], mass spectrometry [5], and DNA detection technology [6] can be used for blood identification. Although these methods have proven reliable and effective for identifying blood and determining its origin, they also have some drawbacks. First and foremost, the above methods extract DNA or protein from cells for analysis in a destructive way. If the amount of sample available is very small, this constitutes a substantial limitation because to the greatest extent possible, blood samples should be protected for further analysis and testing; especially in criminal cases and cases of wildlife protection. In addition, these methods are difficult to apply at the single-cell level. Because hemoglobin differences in red blood cells (RBCs) are the basis for blood differentiation between different species, using individual RBCs for blood identification is more targeted than using whole blood samples. Therefore, exploring a nondestructive, noninvasive, single-cell-based blood identification technology for forensic medicine, veterinary medicine, wildlife conservation, and other related disciplines is of great research value.

#445149 https://doi.org/10.1364/BOE.445149
Journal © 2021 Received 4 Oct 2021; revised 27 Oct 2021; accepted 31 Oct 2021; published 12 Nov 2021

Biomedical Optics EXPRESS

Raman spectroscopy is a powerful detection and analysis technology [7]. Using an inelastic scattering spectrum with different frequencies from incident light, it can obtain information on molecular vibration and rotation, then analyze the structure of the substance. It can therefore be used to determine the "fingerprint" of a substance, and characterize various samples. Compared with other spectrum techniques, Raman spectroscopy has a wide range of detection, including common inorganic substances, organic substances, biological macromolecular synthetic materials (carbon nanotubes), and chemical reaction catalysts, among others [8–12]. It is now widely used in chemistry, materials science, medicine, and other fields [13,14]. It has great value for qualitative analysis, quantitative analysis, and molecular structure determination [15,16]. For the analysis of biological samples, molecular compositional information from cells can be obtained without interference from environmental water, which makes it an ideal tool for biological science research [17].

RBCs are an important component of blood, and much information can be derived from them via Raman spectroscopy. Hemoglobin makes up > 95% of the dry weight of RBCs, and each RBC contains approximately 20 million hemoglobin molecules. The structure of hemoglobin makes the greatest contribution to the molecular vibration profile of RBCs. Structural differences between hemoglobin in different species result in slight differences in their Raman spectrums, which provides a reliable basis for nondestructive and unlabeled blood identification [18]. Previous studies investigating blood classification based on Raman spectroscopy were conducted using whole blood samples [19,20]. If further targeted studies on RBCs are required, Raman spectroscopy should be able to be conducted at the single-cell level.

The main problem with traditional Raman spectroscopy is that the cell samples are in continuous motion due to inherent Brownian motion and the suspension of the aqueous solution, which leads to the failure of stable excitation and acquisition of Raman signals. Laser tweezers Raman spectroscopy (LTRS) combines Optical Tweezers [21] and Raman spectroscopy, which can

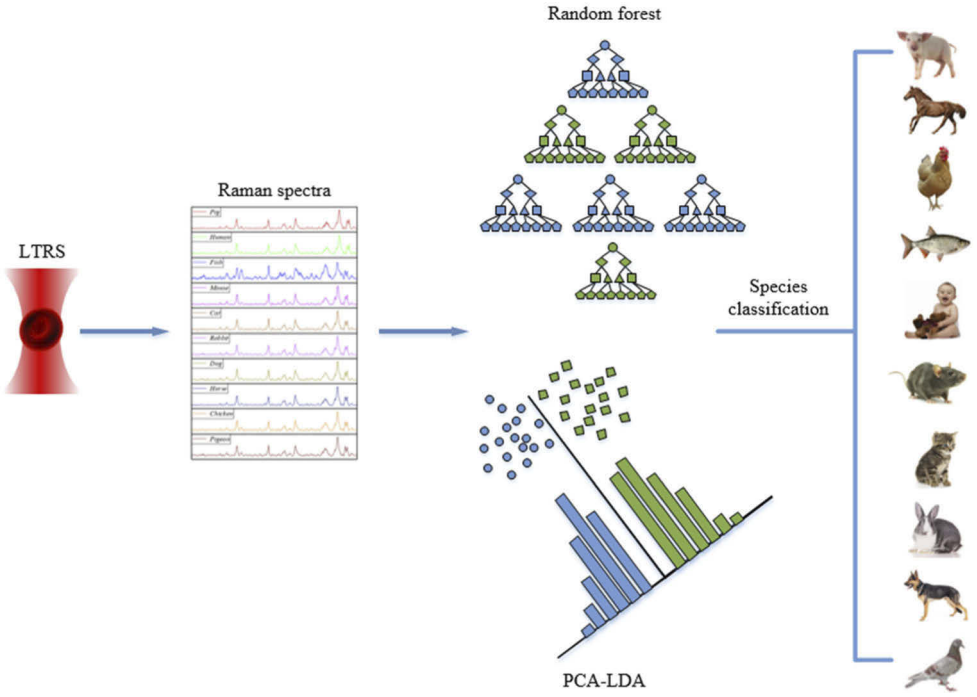

Fig. 1. Experimental flow chart.

Biomedical Optics EXPRESS

effectively solve the difficult problem of detecting the Raman information of particles in a liquid phase environment [22].In Raman detection, optical tweezers are used to capture cells to ensure a single signal source, and the cells can be trapped in a light trap to reduce the influence of the substrate on the background noise of the cell's Raman signal, which is a vast improvement on the traditional method of mechanical fixation of the sample in Raman analysis. With the continuous optimization of machine learning algorithms and data processing methods, LTRS has been widely used in microbiology, biomedicine, environmental science, and other disciplines [23–26].

In the current study the Raman spectra of RBCs from 10 species were obtained without damage or invasion using a self-built LTRS system. After simple preprocessing, two machine learning algorithms were used to analyze spectral data and establish a classification model to predict interspecific blood at the single-cell level; PCA-LDA and random forest (RF). The whole process of this study is shown in Figure 1. In addition, the comparison of the two methods shows that the RF algorithm has a significant advantage that it can directly explain the features that contribute greatly to classification at the molecular level.

2. Materials and methods

2.1. Sample collection and preparation

Anticoagulated whole blood samples from nine non-human species were purchased, including pig(B1612), horse(B1613), dog(B1618), cat(B1617), fish(B1622), chicken(B1614), pigeon(B1623), rabbit(B1616), and mouse(B1620) from Beijing Bersee Science and Technology Co., Ltd. A 2 ml blood sample was extracted from the author's vein using a disposable blood collection needle and stored in a purple vacuum tube containing EDTA anticoagulant for subsequent extraction of RBCs. All samples were centrifuged at 3000 rpm/min for 5 min to remove white blood cells, platelets, plasma, and other impurities. RBCs were then suspended and washed with phosphate-buffered saline three times. Suspensions of RBCs diluted 50 times were generated, then 200 µL of diluted RBC suspension was put into a quartz petri dish which was and placed on the LTRS micromanipulation platform for single RBC detection. In the external environment RBCs are affected by osmotic pressure and temperature, resulting in hemolysis, RBC rupture, and hemoglobin overflow [27]. To prevent RBC hemolysis, all samples were stored in a sterile environment at 4°C–8°C and performed within 24 hours of sample preparation. This research was approved by the Ethics Committee of Beijing Anzhen Hospital, Capital Medical University.

2.2. LTR

Figure 2 shows the structure of the self-made LTRS system and the images of human RBCs before and after capture. The whole set of experimental equipment is consistent with that described in the literature [22]. Briefly, the system consists of a laser, optical coupler, microscope objective, illumination source, charge-coupled device (CCD), spectrometer, photodetector, and computer. The 785-nm near-infrared laser is pumped by a laser diode with adjustable power. First a laser beam of approximately 2 µm is obtained through a beam expansion collimator set, and a narrow bandpass filter is used to purify the output laser wavelength. The laser passes through the oil immersion objective (100; N.A.1.40) and is focused to form an optical potential well of approximately 1 µm for capturing individual RBCs placed on a quartz substrate (approximately 90 µm thick), and the capture laser also stimulates the Raman signal. The back Raman scattering of RBCs is collected by the same objective. The laser is focused on a 600-g/mm grating spectrometer after Rayleigh scattering noise is removed by a notch filter. The recorded spectrum window is 300–2000cm^{-1} and the spectrum resolution is less than 5 cm^{-1}. The Raman signal is split by the grating, and is received by a thermoelectrically-cooled charge-coupled device (1340 × 400 pixels; PIXIS: 400BD). The temperature is reduced to −70°C via thermoelectric cooling, which can effectively reduce the influence of dark current noise on the signal. Spectrum information

Biomedical Optics EXPRESS

is presented on a computer screen for collection and analysis using a software package called WinSpec (Princeton, NJ, USA). The LTRS imaging path is made up of an Olympus microscope and a CCD camera that can facilitate the viewing of captured cells in real time on a computer screen.

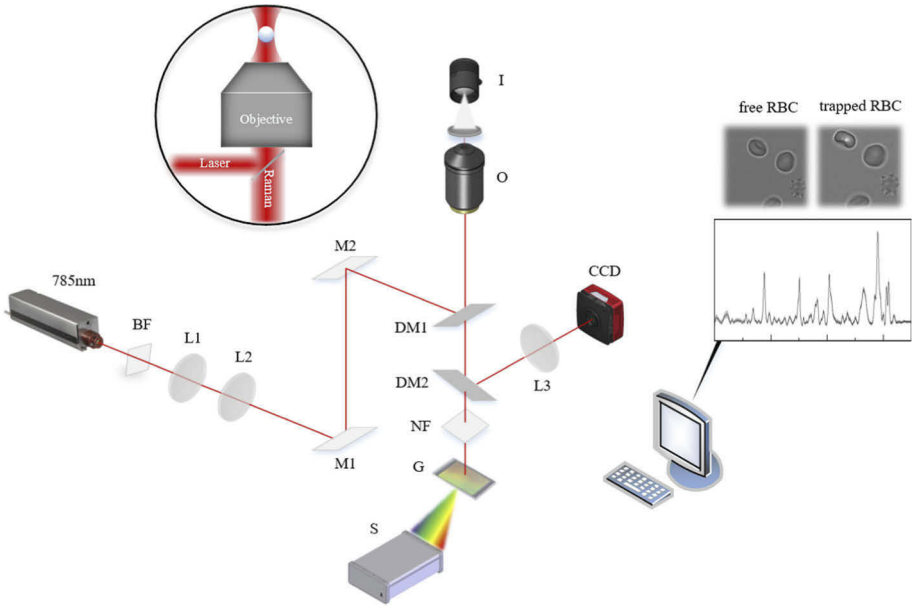

Fig. 2. Schematic diagram of the LTRS system. A 785-nm near-infrared laser serves as both the captured light of the optical tweezers and the excitation light for Raman scattering. Images of captured RBCs and their Raman spectra are displayed on a computer screen. BF, band filter; L, lens; M, mirror; DM, dichroic mirror; O, objective; I, illumination; NF, notch filter; G, grating; S, spectrometer; CCD, charge-coupled device

The selection of light sources was particularly important when building the system. A 785-nm near-infrared laser was selected as the captured light source and Raman excitation light because it can balance the scattering efficiency, photothermal damage, noise, detection efficiency, and other influential factors [28]. In this experiment the average power of the laser was approximately 10 mw, and the exposure time was 30 seconds. Polystyrene spheres (5 μm in diameter) were used for system calibration.

2.3. *Spectra data preprocessing*

Raman spectra of RBCs were obtained at least 450 per species using LTRS, and all these recorded were from different cells. To reduce the influence of intra-individual variation and highlight the common characteristics of specific species, spectra in which cosmic rays were present were removed in advance, and groups of three were averaged to obtain a total of 1500 Raman spectra from 10 species.

When the laser is projected onto the fluorescent material in a sample, the fluorescent background is generated. These background noises affect the acquisition of valuable information in the spectrum, and also cause deviations in the establishment of classification models and prediction results. Therefore, data preprocessing is used to eliminate spectrum data interference by noise before the establishment of the model [29]. The Raman spectra in the range of 400–1700cm^{-1} were intercepted and the quartz substrate background was subtracted. The small error caused by the measuring instrument (*i.e.*, wave number offset caused by different measurement periods) was

Biomedical Optics EXPRESS

corrected. Background fluorescence was removed by baseline calibration using the asymmetric least square method. A Sacitzky-Golay filter with the window size set to 5 was then used for smooth filtering, to suppress high-frequency noise and improve the signal-to-noise ratio of the spectrum. Due to the influence of experimental conditions and laser output fluctuation, it is difficult to compare Raman intensity with different dimensions. To obtain a unified relative Raman intensity and make different spectra comparable, each spectrum was linearly normalized to the range of 0–1. To improve the generalization ability of classification model and avoid the over-fitting phenomenon, two thirds of the preprocessed spectrum data were randomly assigned to a training set, and the remaining third was used as a test set. All the above pretreatment processes and subsequent classification model establishment were performed using the statistical software "Origin 2021" and "R". Figure 3(a) and 3(b) show the average Raman spectra, the images before and after RBCs were captured, and the spectral preprocessing flow chart.

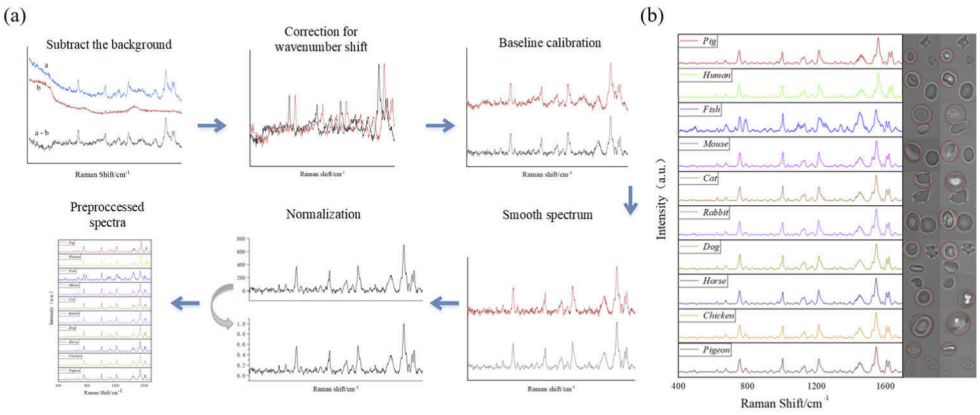

Fig. 3. (a) Flow chart of spectrum data preprocessing. The process consists of subtracting the background generated by the quartz substrate, wavenumber correction, baseline calibration, spectrum smoothing to eliminate noise interference, obtaining relative Raman intensity via spectrum normalization in the range of 0–1, and partition of datasets. (b) Mean Raman spectra of RBCs from 10 species, and images before and after capture (captured cells are circled in red). Solid lines represent mean spectra, and the shadow represents the standard deviation.

3. Results and discussion

3.1. PCA-LDA modeling for RBC classification

PCA in machine learning algorithms is the most widely used unsupervised data dimension reduction algorithm [30]. It maximizes variance by projecting high-dimensional data into low-dimensional data, and then identifies the most important component of the simplified information. Therefore, dimension reduction analysis of PCA fits well with Raman spectroscopy with high dimensional characteristics [31]. The red curve in Fig. 5 shows the cumulative variance of the principal component of the original Raman data after PCA dimensionality reduction. The first two principal components (PCs) have been able to explain more than 70% of the change in all the Raman data. Several spectra are presented in the form of points in a two-dimensional plane or three-dimensional space with PCs as the axis, to achieve the visualization of classification between different cells. Due to a large amount of data, in order to clearly display the three-dimensional distribution of the 10 types of samples, 30 samples were randomly selected from the training set for each species to display the score figure of the first three PCs (PC1, PC2, PC3), as shown in Fig. 4. The spatial distribution of 10 types of samples can be discerned, which also reflects

Biomedical Optics EXPRESS

the advantage of the PCA algorithm; which is that it can present the classification in the most intuitive form.

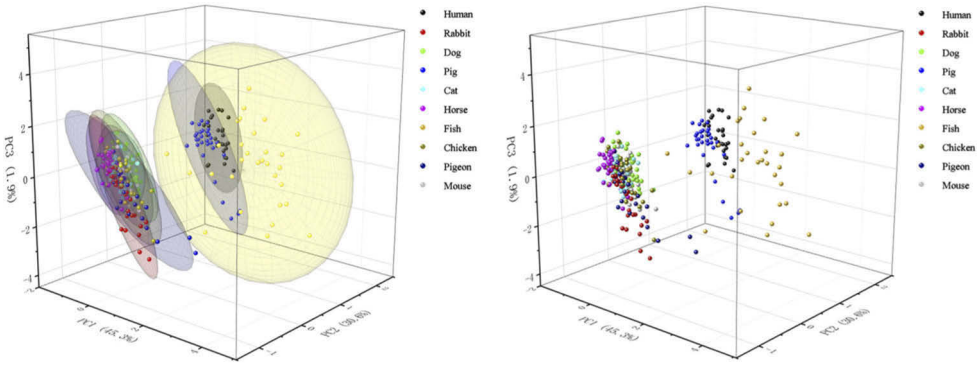

Fig. 4. Three-dimensional visualization of training set classification.

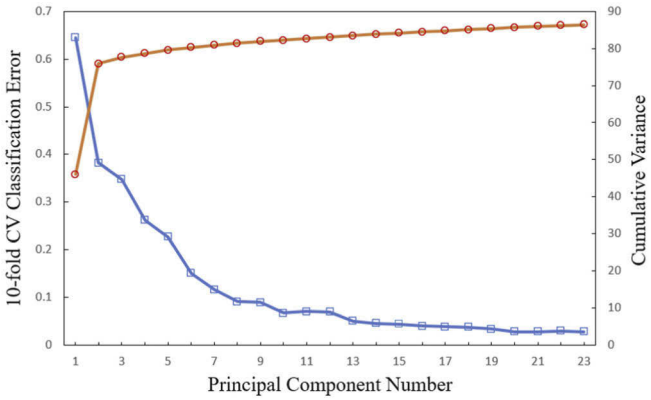

Fig. 5. The blue curve is the relationship between the mean error of 10-fold cross-validation and the number of PCs. The red curve shows the cumulative variance as the number of PCs increases.

Following PCA, Raman spectral data whose original spectral features (wave numbers) are replaced by PCs, so they can be used to build a classification prediction model by a supervised approach such as LDA, which takes advantage of PCA and known classification labels. [32]. In the process of model construction based on PCA-LDA, the most optimal parameter (number of PCs) was selected by using the 10-fold cross-validation method [33]. The whole parameter optimization process involved dividing the training set into 10 groups of datasets of the same size, among which 9 groups of data (90% of the training data) were used for model construction, and the remaining group of data (10% of the training data) was used as a validation set for model evaluation. The selection of validation sets was then changed, and the whole cross-validation process was repeated until all 10 groups of data had been used as validation sets for model evaluation. After each parameter adjustment, the performance of the classification model was evaluated by verifying the average error rate of the set, and the optimal parameter was selected. The blue curve in Fig. 5 shows the process of adjusting parameters. The average error of cross-validation with 20 PCs is the smallest, so the 20 dimensional-reduced features (PC1–PC20) are determined as the optimal parameters. These parameters are used to train the whole training set (100% training data) for the LDA classification model.

Biomedical Optics EXPRESS

The PCA-LDA classification model in the current study had an accurate classification rate of 97.1% for RBCs of 10 species in the training set, and demonstrated the classification ability of the model for different species in the form of a confusion matrix. The classification results are shown in Fig. 6(a). The classification accuracy of human, dog, horse, and pig RBCs is up to 100%. Accuracy was also > 98% for several other species such as cat, rabbit, mouse, and fish in the training model. The two species with the largest classification errors in the entire training model were pigeons (89%) and chickens (87%). In 11 cases pigeons were misclassified as chickens, and in 12 cases chickens were misclassified as pigeons. Chickens and pigeons may have been easily confused in the classification model due to the close relationship between them (*i.e.*, both belong to the avifauna in biological classification).

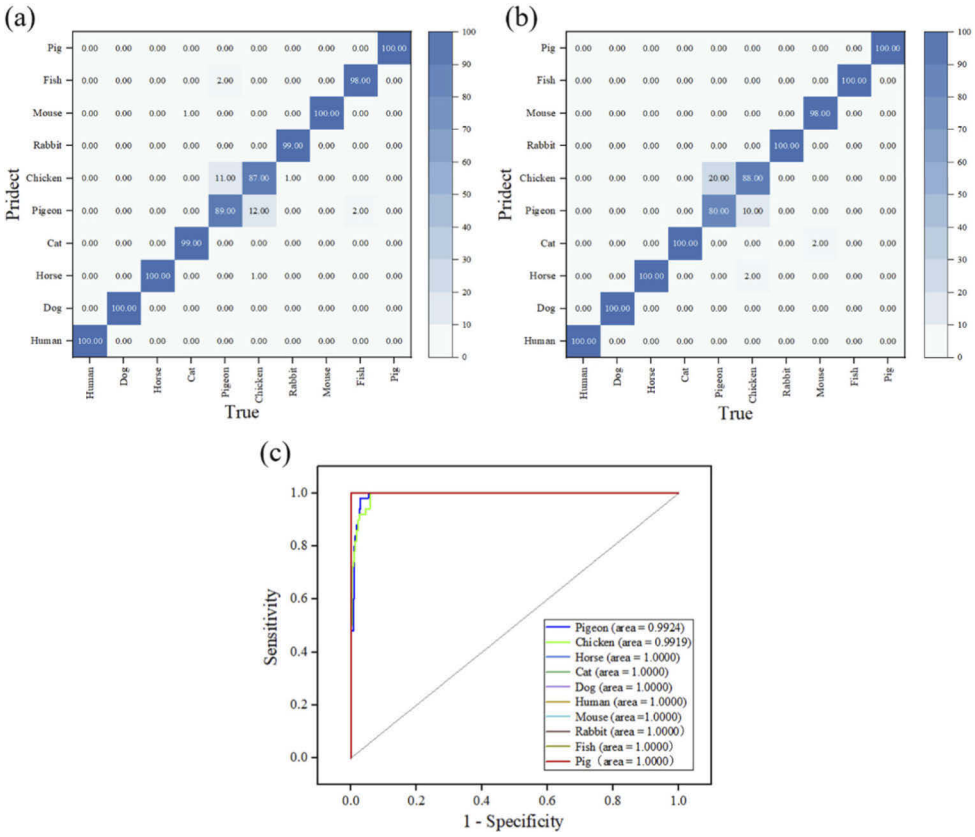

Fig. 6. (a) Prediction results of the PCA-LDA classification model using the training set. (b) Prediction results of the PCA-LDA classification model using the test set. (c) The receiver operating characteristic curve of the test set data in the PCA-LDA classification model indicated that the areas under the curve of all 10 species were greater than 0.99.

To further verify the predictive power of the PCA-LDA classification model, a classification test was conducted using the samples (a total of 500 cases) that were not used in the test set. The spectral data of the test set was loaded into the above model. The results are also presented in the form of a confusion matrix for all correctly classified species; human, dog, horse, cat, rabbit, fish, and pig (Fig. 6(b)). The misjudged species were mouse, pigeon, and chicken, of which only 1/50 cases of mouse classification was incorrect (98% accuracy rate). The classifications of pigeons and chickens were similar those in the training set, with large respective error rates of 80% and

Biomedical Optics EXPRESS

88%. Ten of the pigeon samples were misclassified as chicken samples, and 5 of the chicken samples were misclassified as pigeon samples. The accuracy rate of the whole test set was 96.6%.

To more intuitively reflect the performance of the classification model the receiver operating characteristic (ROC) curve was used [34], which describes the relationship between the sensitivity of prediction and the false-positive rate at different thresholds. The area under the ROC curve (AUC; range 0–1) is positively correlated with prediction accuracy. The closer the AUC is to 1, the more ideal the prediction model is. In the present study the AUCs of all species were greater than 0.99 (Fig. 6(c)), indicating that the combination of LTRS and PCA-LDA had high specificity and sensitivity, and could achieve high-precision identification of blood species origin at the single-cell level.

3.2. RF modeling for RBC classification

RF is a machine learning algorithm that synthesizes multiple decision trees via ensemble learning, and generates predictions based on numerous factors [35]. Its basic unit is a decision tree. Intuitively speaking, every decision tree is a weak classifier. When a prediction variable is entered, N decision trees will produce N classification results. RF integrates the prediction results of all tree voting, and designates the prediction category with the most votes the final output. Rules generated by the RF are based on the bootstrapping method [36], and the same number of samples are randomly selected and put back to "train" each decision tree. This method can effectively avoid the defect of high variance associated with a single decision tree, to achieve better classification performance. As well as high prediction accuracy, RF has many advantages [37,38] such as the capacities for efficient use with large datasets, processing samples with high dimensional features without dimensionality reduction, and assessing the importance of each feature in classification problems, among others. These advantages are highly compatible with single-cell Raman spectroscopy, so RF was also used to build a classification model to identify RBCs of different species in the current study.

The performance of an RF classification model depends on two important parameters; the number of characteristic variables used by each decision tree (M), and the number of trees in the "forest" (N). Increases in M and N improve the statistical variance of the model, reduce prediction error, and increase computational complexity and time, so it is necessary to select an optimal parameter. An advantage of RF is that parameter optimization can be conducted without cross-validation. The characteristic data obtained via bootstrapping is a subset of the training set, and the remaining observation values are referred to as "out-of-bag" (OOB). The overall classification error can be calculated by using the OOB error. When the amount of data is large enough, the OOB error is essentially equivalent to the cross-validation error. Figure 7 shows the relationship between the two parameters and the OBB error. When $M = 30$ and $N = 600$, the OOB error is stable at 4.4% which is also the optimal value to balance calculation efficiency and prediction accuracy. Because RF is a classification model of random sampling, each decision tree forming the forest is different, which also leads to different prediction results each time. Therefore, the OOB error used here refers to the average error after multiple calculations.

A RF classifier was then constructed by determining the best parameters and using training set data (1000 cases without PCA dimensionality reduction). This RF classifier was used with the test set to evaluate the performance of the model. Due to the randomness of the RF, the model was repeatedly run on the training set and test set 10 times, and the average respective accuracies of the classification were 96.62% and 96.84% (Table 1). The classification results of a randomly selected sixth test set (accuracy 96.8%) are presented as a confusion matrix. The ROC was derived from the predicted probability of 10 species in the test set to evaluate the performance of the classification model. The AUCs of all species were greater than 0.98 (Fig. 8(b)), which also indicated that the RF classification model had high specificity and sensitivity.

Biomedical Optics EXPRESS

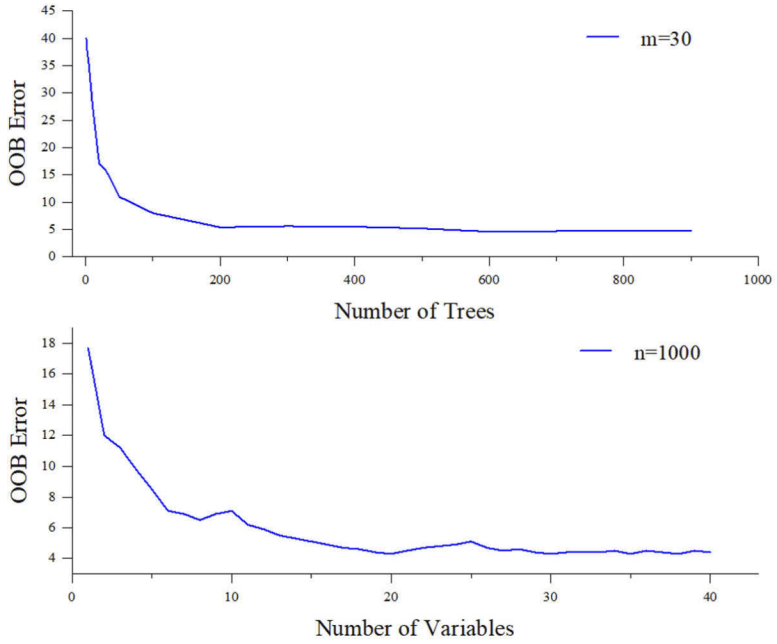

Fig. 7. Relationships between out-of-bag error and two important parameters in RF. N is the number of decision trees in the forest, and M is the number of selected features in each decision tree.

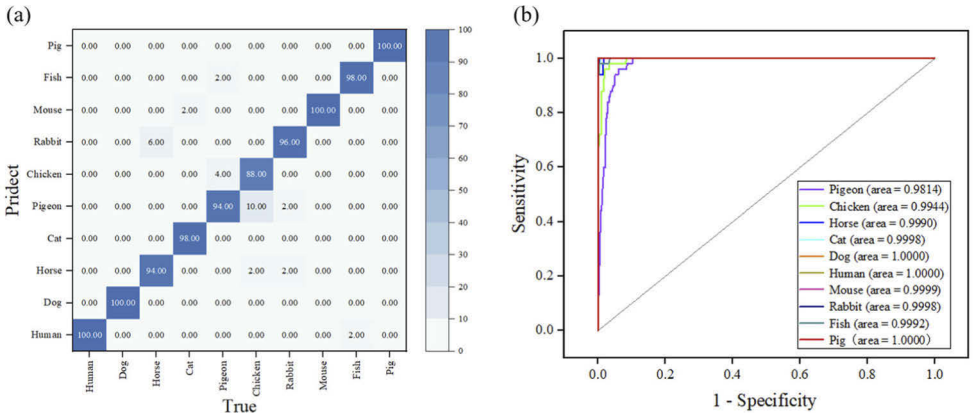

Fig. 8. (a) Prediction results of the RF classification model using the test set. (b) The ROC curve of the test set data in the RF classification model showed that the AUCs of all 10 species were greater than 0.98.

Table 1. Average accuracies of the RF classification model run 10 times in the training set and the test set.

Phase	1	2	3	4	5	6	7	8	9	10	Average
Training	96.4	96.8	97.2	96.4	97	96	96.4	97.4	96	96.6	96.62
Test	97	96.2	97	96.8	97.4	96.8	97	97	97	96.2	96.84

Biomedical Optics EXPRESS

3.3. Advantage of RF modeling

In terms of classification accuracy, the two machine learning algorithms have similar classification performance. But how to explain which features (wavenumbers) contribute the most to classification? Because the PCA-LDA somehow transforms the original Raman features into PCs, although it achieves effective classification, it loses the interpretability of important features. In previous studies partial least-squares-discriminant analysis [39], support vector machine [40], convolutional neural network [22] was unable to explain which features contributed the most to the classification process at the molecular level, thus failing to discover potential special Raman peak locations. The most significant advantage of the RF algorithm is that it can directly extract the features that contribute most to classification. Unlike the features (PCs) in PCA-LDA, RF is the analysis of no dimensionally reduced features, so the Gini index [41] can be used to describe the importance of variables in the Raman spectrum, which can be understood as an indicator to measure the purity of nodes in the decision tree (*i.e.*, if the Gini index is small it means that almost all variables in a node came from the same category). Figure 9 shows the top 22 most important features extracted from the classification results of the test set data, which are the features that reduce the Gini index the most. Bands 1550–1554 cm^{-1}, 1566–1568 cm^{-1} related to heme in hemoglobin, and 1003–1006 cm^{-1} related to phenylalanine contribute the most to classification prediction [42].

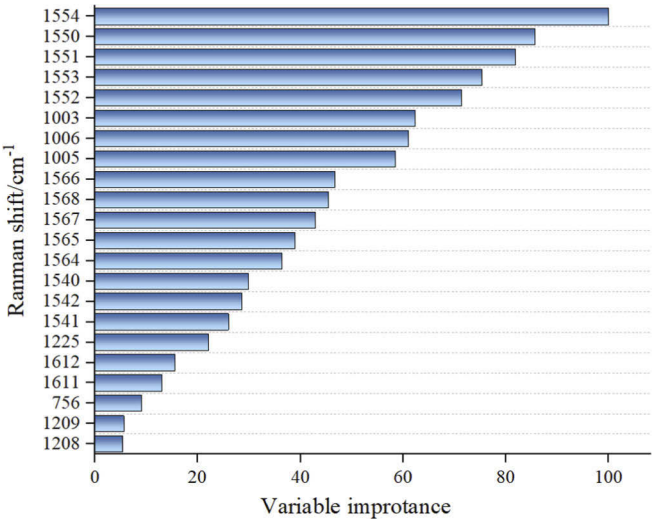

Fig. 9. Variable importance description of the test set. Variable importance is calculated from the average decrease in the Gini coefficient and expressed as the relative value of the maximum and minimum.

Raman spectrum peak position analysis can be used to acquire molecular structure information of various functional groups and chemical bonds, determine potentials and most different Raman characteristic peaks, and distinguish the Raman spectrum representing different samples based on these parameters. Hemoglobin is the main component of RBCs, and its highly conjugated heme subunit has a relatively large Raman cross-section, so the Raman spectrum of hemoglobin mainly reflects the vibration characteristics of heme [43]. Therefore, the Raman peaks of RBCs and whole blood samples are related to hemoglobin and its component heme. Previous studies have also shown that protein abundance is a critical factor in the classification of different cells [44].

To avoid the small error caused by the measurement instrument, samples of horse and dog measured on the same date were selected for comparison. Figure 10(a) shows the average Raman spectra of 100 horse erythrocytes (red curve) and 100 dog erythrocytes (black curve) pretreated

Biomedical Optics EXPRESS

in the training set. The shape of the Raman spectra of RBCs in the two different species is very similar because the cells have the same biomolecular structure. The mean spectrum at 621, 673, 754, 1003, 1128, 1212, 1545, 1606, and 1620 cm^{-1} shared common characteristics among different RBCs. The Raman peaks of these wave numbers were derived from the vibration of hemoglobin and other molecules related to hemoglobin. The blue curve in Fig. 10(a) represents the difference between the two mean spectral lines, which is consistent with the importance of variables described in Fig. 9 by observing several bands with large differences. But we found that the variables of importance provided in Fig. 9, do not match up with the Raman band positions seen in the spectra (as given in Grey shaded areas of Fig. 10(a)), instead there is a small shift in position between the variable of interest and the Raman band position. The reason for this phenomenon is that the Raman peaks of different species are misplaced due to the spectral pre-processing step of wavenumber shift correction before the classification model is constructed. Such as 1550 cm^{-1} and 1568 cm^{-1} on the sides of 1545 cm^{-1}, 1225 cm^{-1} and 1209 cm^{-1} on the sides of 1212 cm^{-1}, 756 cm^{-1} on the right of 754 cm^{-1}, 1612 cm^{-1} on the right of 1606 cm^{-1}, and 1005 cm^{-1} on the right of 1003 cm^{-1}. These specific distributions of peaks are shown in Table 2 [45–47] (including the wave number of the offset). In addition, the differences of ten types of RBCs at 1554cm^{-1} and 1003cm^{-1} were quantitatively analyzed in the form of the boxplot, as shown in Fig. 10(b) and 10(c). The bands with significant differences such as these contribute significantly to cell classification and were consistent with the analysis in Fig. 9. Therefore, compared with other algorithms (including PCA-LDA), the RF algorithm is proved can achieve

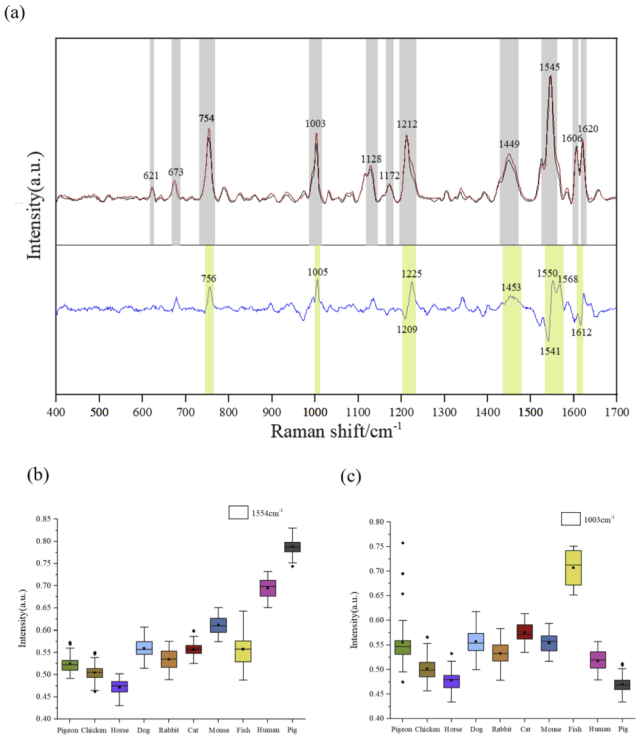

Fig. 10. (a) Comparison of preprocessed Raman spectra of horse RBCs (red curve) and dog RBCs (black curve). Grey shaded areas represent Raman characteristic peaks. The blue curve is the difference spectrum (human minus pig). Green shaded areas represent Raman characteristic peaks with great difference. (b) Plot of differences in intensity between 10 species at 1554 cm^{-1}. (c) Plot of differences in intensity between 10 species at 1003 cm^{-1}.

Biomedical Optics EXPRESS

accurate classification while directly obtaining important spectral features that contribute to classification.

Table 2. Raman shift assignment of RBCs spectra.

Raman shift (cm^{-1})	Assignment[a]	Component
621	C–C twist	Phe
673	δ (pyr deform) $_{sym}$	Trp
754 (756)	ν15, ν (pyr breathing)	Trp
1003 (1005)	phenylalanine	Phe
1128	ν (C_β-methyl)	glucose
1172	ν (pyr half-ring) $_{asym}$	Trp
1212 (1209)	ν5 + ν18, δ (C_mH)	protein
1225	..(C_mH)	protein
1449 (1453)	δ (CH_2/CH_3)	protein
1545 (1550)	ν11, ν ($C_\beta C_\beta$)	heme
1568	ν19, ν ($C_\beta C_\beta$)	heme
1606(1612)	ν (C = C) $_{vinyl}$	heme
1620	ν (C = C) $_{vinyl}$	heme

[a]Assignments taken from Refs. [45–47] ν, stretch; δ, bend/scissor; sym, symmetric; p, protein; pyr, porphyrin; Phe, phenylalanine; Trp, tryptophan.

4. Conclusion

Using a self-built LTRS system, we obtained RBC Raman spectra of 10 species without destroying the erythrocyte. Two machine learning algorithms (PCA-LDA and RF) were used for the statistical processing of spectral data and the establishment of a classification model. The prediction accuracy of the two algorithms for 10 types of RBCs was > 96%. Furthermore, the advantages of RF algorithm compared with PCA-LDA and other classification algorithms are analyzed, which provides a reference for spectroscopist to select targeted classification algorithms in the future.

We demonstrated that LTRS combined with machine learning algorithms is an effective analytical technique for interspecific blood classification at the single-cell level. Optical tweezers can selectively separate a large number of samples in the liquid environment, and the excited Raman spectrum can be targeted for the analysis of a single micro-nano sample. LTRS has the advantages of being nondestructive and noninvasive, which makes it ideal for studying numerous types of biological samples, especially when the sample size is limited and destructive modes of identification are not possible. This is of great significance in the fields of clinical and rare species research. Therefore, combining more advanced artificial intelligence algorithms to achieve higher accuracy of biological cell identification is the main direction of future development. LTRS combined with machine learning algorithms can be a useful tool for scientists exploring the biological world at the molecular level.

Funding. Beijing Youth Talent Support Program (Z2019042); Beijing Great Wall Scholars Program (CIT& TCD20190323); National Natural Science Foundation of China (61875237, 81900504).

Acknowledgments. It was appreciated that professors Guixian Zhu and Qiang Yang gave some very useful suggestions for preparation of the manuscript. We thank Dr. Zhao from Beijing Anzhen Hospital, Capital Medical University for his guidance in blood collection.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Biomedical Optics EXPRESS

References

1. A. E. Levin, P. C. Williamson, E. M. Bloch, J. Clifford, S. Cyrus, B. H. Shaz, D. Kessler, J. Gorlin, J. L. Erwin, N. X. Krueger, G. V. Williams, O. Penezina, S. R. Telford, J. A. Branda, P. J. Krause, G. P. Wormser, A. M. Schotthoefer, T. R. Fritsche, and M. P. Busch, "Serologic screening of United States blood donors for Babesia microti using an investigational enzyme immunoassay," Transfusion **56**(7), 1866–1874 (2016).
2. D. V. Sotnikov, A. V. Zherdev, and B. B. Dzantiev, "Mathematical model of serodiagnostic immunochromatographic assay," Anal. Chem. **89**(8), 4419–4427 (2017).
3. S. S. Tobe, N. Watson, and N. N. Daéid, "Mathematical model of serodiagnostic immunochromatographic assay," J. Forensic Sci. **52**(1), 102–109 (2007).
4. L. Dobrila, T. Zhu, D. Zamfir, M. Tarnawski, R. Ciubotariu, M. S. Albano, M. DeLeon, K. Turner, M. Azimi, C. C. Hoppe, A. Scaradavou, and P. Rubinstein, "Detection of hemoglobin (Hb) variants by HPLC screening in cord blood units (CBU) donated to the national cord blood program (NCBP)," Blood **128**(22), 2182 (2016).
5. Y. R. Xie, D. C. Castro, S. E. Bell, S. S. Rubakhin, and J. V. Sweedler, "Single-cell classification using mass spectrometry through interpretable machine learning," Anal. Chem. **92**(13), 9338–9347 (2020).
6. M. X. Chen, L. Ai, J. H. Chen, X. Y. Feng, S. H. Chen, Y. C. Cai, Y. Lu, X. N. Zhou, J. X. Chen, and W. Hu, "DNA microarray detection of 18 important human blood protozoan species," PLoS Negl Trop Dis **10**(12), e0005160 (2016).
7. G. W. Auner, S. K. Koya, C. Huang, B. Broadbent, M. Trexler, Z. Auner, A. Elias, K. C. Mehne, and M. A. Brusatori, "Applications of Raman spectroscopy in cancer diagnosis," Cancer Metastasis Rev. **37**(4), 691–717 (2018).
8. Y. T. Gong, B. H. Li, T. Pei, C. H. Lin, and S. Lee, "Raman investigation on carbonization process of metal–organic frameworks," J. Raman Spectrosc. **47**(10), 1271–1275 (2016).
9. N. Wattanavichean, M. Gilby, R. J. Nichols, and H. Arnolds, "Detection of metal–molecule–metal junction formation by surface enhanced Raman spectroscopy," Anal. Chem. **91**(4), 2644–2651 (2019).
10. M. Jia, S. Li, L. Zang, X. Lu, and H. Zhang, "Analysis of biomolecules based on the surface enhanced Raman spectroscopy," Nanomaterials **8**(9), 730 (2018).
11. S. Chaunchaiyakul, T. Yano, K. Khoklang, P. Krukowski, M. Akai-Kasaya, A. Saito, and Y. Kuwahara, "Nanoscale analysis of multiwalled carbon nanotube by tip-enhanced Raman spectroscopy," Carbon **99**, 642–648 (2016).
12. C. Hess, "New advances in using Raman spectroscopy for the characterization of catalysts and catalytic reactions," Chem. Soc. Rev. **50**(5), 3519–3564 (2021).
13. S. Y. Ding, J. Yi, J. F. Li, B. Ren, D. Y. Wu, R. Panneerselvam, and Z. Q. Tian, "Nanostructure-based plasmon-enhanced Raman spectroscopy for surface analysis of materials," Nat. Rev. Mater. **1**(6), 16021 (2016).
14. G. Pezzotti, "Raman spectroscopy of biomedical polyethylenes," Acta Biomater. **55**, 28–99 (2017).
15. P. Bhosale, I. V. Ermakov, M. R. Ermakova, W. Gellermann, and P. S. Bernstein, "Resonance Raman quantification of nutritionally important carotenoids in fruits, vegetables, and their juices in comparison to high-pressure liquid chromatography analysis," J. Agric. Food Chem. **52**(11), 3281–3285 (2004).
16. M. O'Connell, A. G. Ryder, M. N. Leger, and T. Howley, "Qualitative analysis using Raman spectroscopy and chemometrics: a comprehensive model system for narcotics analysis," Appl. Spectrosc. **64**(10), 1109–1121 (2010).
17. A. Pitters, M. Cuellar, P. Wiegand, S. Gilliam, I. R. Lewis, and B. Lenain, "Raman spectroscopy as a real-time in situ analyzer for cell culture bioprocesses," New Biotechnol. **33**, S113 (2016).
18. C. G. Atkins, K. Buckley, M. W. Blades, and R. F. B. Turner, "Raman spectroscopy of blood and blood components," Appl. Spectrosc. **71**(5), 767–793 (2017).
19. G. McLaughlin, K. C. Doty, and I. K. Lednev, "Raman spectroscopy of blood for species identification," Anal. Chem. **86**(23), 11628–11633 (2014).
20. K. Virkler and I. K. Lednev, "Blood species identification for forensic purposes using Raman spectroscopy combined with advanced statistical analysis," Anal. Chem. **81**(18), 7773–7777 (2009).
21. J. L. Killian, F. Ye, and M. D. Wang, "Optical tweezers: a force to be reckoned with," Cell **175**(6), 1445–1448 (2018).
22. W. L. Lu, X. Q. Chen, L. Wang, H. F. Li, and Y. V. Fu, "The combination of an artificial intelligence approach and laser tweezers Raman spectroscopy for microbial identification," Anal. Chem. **92**(9), 6288–6296 (2020).
23. D. Lin, Z. Zheng, Q. Wang, H. Huang, Z. Huang, Y. Yu, S. Qiu, C. Wen, M. Cheng, and S. Feng, "Label-free optical sensor based on red blood cells laser tweezers Raman spectroscopy analysis for ABO blood typing," Opt. Express **24**(21), 24750–24759 (2016).
24. S. Qiu, M. Li, J. Liu, X. Chen, T. Lin, Y. Xu, Y. Chen, Y. Weng, Y. Pan, S. Feng, X. Lin, L. Zhang, and D. Lin, "Study on the chemodrug-induced effect in nasopharyngeal carcinoma cells using laser tweezer Raman spectroscopy," Biomed. Opt. Express **11**(4), 1819–1833 (2020).
25. S. P. Zhang, Y. C. Sun, B. B. Liu, and R. Li, "Full size microplastics in crab and fish collected from the mangrove wetland of Beibu Gulf: evidences from Raman tweezers (1–20 μm) and spectroscopy (20–5000 μm)," Sci. Total Environ. **759**, 143504 (2021).
26. W. E. Huang, A. D. Ward, and A. S. Whiteley, "Raman tweezers sorting of single microbial cells," Environ. MicroBiol. **1**(1), 44–49 (2009).
27. S. Streichman, E. Kahana, and I. Tatarsky, "Hypertonic cryohemolysis of pathologic red blood cells," Am. J. Hematol. **20**(4), 373–381 (1985).
28. C. N. LaFratta, "Optical tweezers for medical diagnostics," Am. J. Hematol. **405**, 5671–5677 (2013).

Biomedical Optics EXPRESS

29. N. K. Afseth, V. H. Segtnan, and J. P. Wold, "Raman spectra of biological samples: a study of preprocessing methods," App. Spectrosc. **60**(12), 1358–1367 (2006).
30. L. J. Li, S. G. Liu, Y. L. Peng, and Z. G. Gou, "Overview of principal component analysis algorithm," Optik **127**(9), 3935–3944 (2016).
31. X. Y. Cui, Z. Y. Zhao, G. J. Zhang, S. Chen, Y. Zhao, and J. Lu, "Analysis and classification of kidney stones based on Raman spectroscopy," Biomed. Opt. Express **9**(9), 4175–4183 (2018).
32. J. Y. Lin, L. D. Shao, S. F. Qiu, X. W. Huang, M. M. Liu, Z. C. Zheng, D. Lin, Y. L. Xu, Z. H. Li, Y. Lin, R. Chen, and S. Y. Feng, "Application of a near-infrared laser tweezers Raman spectroscopy system for label-free analysis and differentiation of diabetic red blood cells," Biomed. Opt. Express **9**(3), 984–993 (2018).
33. T. T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," Pattern Recogn **48**(9), 2839–2846 (2015).
34. N. A. Obuchowski, "Receiver operating characteristic curves and their use in radiology," Radiology **229**(1), 3–8 (2003).
35. R. Genuer, J. M. Poggi, C. Tuleau-Malot, and N. Villa-Vialaneix, "Random forests for big data," Big Data Res **9**, 28–46 (2017).
36. E. Scornet, "Random forests and kernel methods," IEEE Trans. Inform. Theory **62**(3), 1485–1500 (2016).
37. Y. Ao, H. Q. Li, L. P. Zhu, A. Ali, and Z. G. Yang, "The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling," J. Pet. Sci. Eng **174**, 776–789 (2019).
38. Q. Wang, T. T. Nguyen, J. Z. Huang, and T. T. Nguyen, "An efficient random forests algorithm for high dimensional data classification," Adv. Data Anal. Classif **12**(4), 953–972 (2018).
39. L. A. Bratchenko, L. A. Bratchenko, A. A. Lykina, M. V. Komarova, D. N. Artemyev, O. O. Myakinin, A. A. Moryatov, L. L. Davydkin, S. V. Kozlov, and V. P. Zakharov, "Comparative study of multivariative analysis methods of blood Raman spectra classification," J. Raman Spectrosc **51**(2), 279–292 (2020).
40. C. G. Zheng, S. Qing, J. Wang, G. D. Lü, H. Y. Li, X. Y. Lü, C. L. Ma, J. Tang, and X. X. Yue, "Diagnosis of cervical squamous cell carcinoma and cervical adenocarcinoma based on Raman spectroscopy and support vector machine," Photodiagn. Photodyn. Ther **27**, 156–161 (2019).
41. S. Nembrini, "On what to permute in test-based approaches for variable importance measures in random forests," Bioinformatics **35**(15), 2701–2705 (2019).
42. A. Bankapur, E. Zachariah, S. Chidangil, M. Valiathan, and D. Mathur, "Raman tweezers spectroscopy of live, single red and white blood cells," PLoS ONE **5**(4), e10427 (2010).
43. T. G. Spiro and T. C. Strekas, "Resonance Raman spectra of hemoglobin and cytochrome c: inverse polarization and vibronic scattering," Proc. Natl. Acad. Sci. U.S.A. 69(9), 2622–2626 (1972).
44. P. Bai, J. Wang, H. Yin, Y. Tian, W. Yao, and J. Gao, "Discrimination of human and nonhuman blood by Raman spectroscopy and partial least squares discriminant analysis," Anal. Lett. **50**(2), 379–388 (2017).
45. B. R. Wood, P. Caspers, G. J. Puppels, S. P. andiancherri, and D. McNaughton, "Resonance Raman spectroscopy of red blood cells using near-infrared laser excitation," Anal. Bioanal. Chem **387**(5), 1691–1703 (2007).
46. S. Barkur, A. Bankapur, S. Chidangil, and D. Mathur, "Effect of infrared light on live blood cells: role of β-carotene," J. Photochem. Photobiol. B Biol **171**, 104–116 (2017).
47. E. Zachariah, A. Bankapur, C. Santhosh, M. Valiathan, and D. Mathur, "Probing oxidative stress in single erythrocytes with Raman tweezers," J. Photochem. Photobiol. B Biol **100**(3), 113–116 (2010).